Prediction of protein subunits using KEGG BRITE

Oscar Dias, Filipe Liu, Miguel Rocha, Eugénio Ferreira, Isabel Rocha

CEB - Centre of Biological Engineering, University of Minho, 4710-057 Braga, Portugal

The increased importance of genome-scale metabolic models (GSMMs) within systems biology and metabolic engineering, led to the development of several computational frameworks dedicated to their reconstruction. One of the toughest challenges, when reconstructing a model is associated to the identification of gene-protein-reaction (GPR) associations, a step usually performed by manually searching literature. In this work, we present a new approach for automatically predicting, at the genome level, protein subunits using the KEGG BRITE database. This database contains information on hundreds of protein complexes, which can be automatically retrieved using the KEGG representational state transfer (REST) application programing interface (API). Afterwards, the gene association rule related to each protein complex is individually processed by running it through a grammar specially developed to parse these data. The parsed rule is then fitted to the genome annotation, to determine if the complex is encoded in the case study genome. Finally, the GP rule can be integrated into a metabolic model to formulate a GPR association. This methodology is implemented and can be automatically performed in *merlin*, a user-friendly Java application that performs the reconstruction of genome-scale metabolic models previously developed by the authors.